Dramatic differences in global dynamic properties and activity of closely-related consensus-designed variants of triosephosphate isomerase

Brandon J. Sullivan[1]

ABSTRACT

Consensus design, meaning the selection of mutations based on the most common amino acid in each position of a multiple sequence alignment, has proven to be an efficient way to engineer stabilized mutants and even to design entire proteins. However, its application has mainly been limited to small motifs or small families of related proteins. Also, we have little idea of how information that specifies a protein's properties is distributed between positional effects (consensus) and interactions between positions (correlated occurrences of amino acids). Here, we designed several consensus variants of triosephosphate isomerase, a large, diverse family of complex enzymes. The first variant was only weakly active, had molten globular characteristics, and was monomeric at 25 ºC despite being based on nearly all dimeric enzymes. A closely related variant that resulted from curation of the sequence database resulted in a native-like, dimeric TIM with near diffusion-controlled kinetics like the wild-type enzyme. Both of these enzymes vary substantially (30-40%) from any natural TIM, but they vary from each other in only a small number of unconserved positions. We demonstrate that sufficient information is contained in the consensus sequence to engineer a sophisticated protein that requires precise substrate positioning and coordinated loop motion. We show that the difference in oligomeric states and native-like properties for the two consensus variants is not a result of defects in the dimerization interface, but rather disparate global dynamic properties of the proteins. These results also have important implications for the role of correlated occurrences of amino acids in

[1] Author would like to thank coauthors Venuka Durani and Thomas J. Magliery

specifying protein properties, the ability of TIM to function as a monomer, and the ability of molten globular proteins to carry out complex reactions.

INTRODUCTION

The sequence of amino acids in a protein encodes its physical and functional properties, but our ability to read that code is still very limited (1). For example, there have been great successes in computational prediction and design of proteins in recent years (2, 3), but we are still far from a comprehensive, accurate model of the thermodynamic consequences of mutations (4, 5). In part this is because natural proteins are typically only stabilized by 5-15 kcal mol^{-1} over the unfolded state, and our knowledge of how to model the unfolded state is poor (6, 7). Remarkable functional designs of enzymes have also been achieved recently, but it remains exceedingly difficult to achieve catalytic efficiencies that compare to natural enzymes (8-10). The effects of solvation, backbone motion, dynamics and entropy are largely beyond our ability to predict or design.

One method of designing non-natural sequences with native-like structures and functions is to look to statistical analysis of families of natural proteins. Genomic sequencing has given us vast databases of sequences of proteins that all have approximately the same structure and activity. This is basically a post-genomic formulation of the so-called "inverse folding problem": what are all sequences in nature that adopt a particular fold (11)? In the limit, the conservation and variation of sequence features in a multiple sequence alignment (MSA) must contain all of the information necessary to design stable, active sequences. The question is: how do we read and apply that information? We were particularly interested in determining what information is encoded at the positional level (consensus/conservation) versus what is encoded by coupling between sites (correlation).

The idea of designing proteins, domains or motifs from consensus is attractive because it makes intuitive sense that the most common amino acid in each position of a multiple sequence alignment is there for a reason (structural, functional, dynamic, etc.). Consensus sequences of motifs like the tetratricopeptide repeat (TPR) and ankyrin repeat, and even small families of related proteins like the fungal phytases, have been shown to be folded (12-17). These consensus-designed proteins generally have higher thermal stabilities than the average proteins from which the consensus sequence was derived. (However, some rational design considerations were applied to unconserved sites in many of these studies.) Data from antibodies and thioredoxin suggest that about half the time, mutation of an amino acid to the most common amino acid in the MSA for that position is stabilizing (18-22).

On the other hand, the most common amino acid in an unconserved site presumably has little informational value, and furthermore unconserved sites may be correlated to each other, which is lost in the consensus. For example, the consensus sequence of TPR motifs has a canonical charge of -7 even though individual TPRs have a 0 ± 2.5 net charge, because the charged residues are largely poorly conserved surface residues which exhibit charge neutralization only when correlation is considered (23). The distribution of information between consensus and correlation is not known, although design of WW domains using only consensus versus consensus plus correlation yielded a much larger fraction of folded proteins with incorporation of the correlation data (24, 25). When triosephosphate isomerase was extensively mutated, virtually all structural positions could individually be mutated conservatively (e.g, Gln to Asn) with little effect on activity, but when all positions were simultaneously varied between the natural residue and a conservative replacement, only about 1 in 10^{10} was active (26). Therefore, interactions among sites appear to account for a great deal of the information in specifying a

folded, active protein, but no experiments to date have elucidated the exact effects of correlated mutations.

To start to answer this question, we proposed to engineer the pure consensus sequence of a complex protein architecture from a large, diverse enzyme family. Presumably, this pure consensus sequence would scramble or ablate many of the sequence correlations at poorly conserved sites, and as such could act as 'host' for interrogating the effects of 'guest' correlation mutations. We selected the triosephosphate isomerases (TIMs) for this study, because they are a very-well studied archtypical member of the (□/□)$_8$ proteins that make up 10% of all biological catalysts (27-29). Because of their glycolytic function in the isomerization of dihydroxyacetone phosphate (DHAP) and glyceraldehyde-3-phosphate (GAP), virtually every organism has a TIM and therefore hundreds of sequences are available. TIM catalyzes a sophisticated reaction with nearly diffusion-limited kinetics and with coordinated motion in the catalytic cycle (30-35). Furthermore, TIM barrel proteins have generally been difficult to engineer despite their ubiquity in nature (36).

Here we report the construction and characterization of two closely-related TIM proteins based purely on consensus, one from a "raw" sequence database and one from a later database curated of fragments and repeats. The raw consensus TIM (cTIM) is weakly active, poorly folded and monomeric, in contrast to nearly all known natural TIMs, which are dimers. The curated consensus TIM (ccTIM) is dimeric, well-folded and fully active. We demonstrate that the oligomeric states are not a result of mutations at the interface, but rather that global dynamic properties of the proteins differ dramatically. Those properties arise from sequence variations at unconserved sites, where correlated sequence networks may play a significant role.

MATERIALS AND METHODS

See *SI* for detailed materials and experimental procedures.

Generation of TIM variants. cTIM was designed from the Pfam HMM multiple sequence alignment of 639 TIMs (37). The most common amino acid was selected at each highly-occupied position to yield a TIM with the same number of amino acids (248) as yeast TIM. The ccTIM sequence was constructed similarly from an updated Pfam dataset that was curated of sequence fragments and repeated sequences. The cTIM, ir-cTIM (*vide infra*), and ccTIM genes were constructed from PCR assembly of overlapping synthetic oligonucleotides. The *S. cerevisiae* TIM gene was cloned from the genome of yeast strain YPH499.

Proteins. Constructs were cloned and expressed as TEV protease-cleavable N-terminal 6×His-fusions under the control of the T7 promoter. TIM variants were purified from BL21(DE3) and a DE3-lysogenized TIM knockout from the Keio collection (38) by Ni-NTA affinity chromatography. The 6×His tag was removed by TEV protease digestion and a second Ni-NTA chromatography step.

Activity. Michaelis-Menten parameters were determined using the coupled assays described first by Plaut & Knowles and applied by the Richards Lab (39, 40). *In vivo* activity was assayed by complementing a TIM-knockout Keio(DE3) strain on minimal media containing glycerol or lactate.

Folding and Dynamics. Ellipticity was recorded on a Jasco J-815 Circular Dichroism spectrometer at 12 μM. 1,8-ANS binding and fluorescence was recorded on a Perkin-Elmer LS50B fluorimeter with 5 μM dye and 25 μM protein. $^{1}H,^{15}N$-HSQC NMR was performed on a Bruker DRX 600 MHz NMR at 350 μM protein.

Oligomeric State. Gel filtration chromatography was performed on a Pharmacia FPLC at 37 μM protein. Sedimentation velocity was performed at the University of Connecticut's Analytical Ultracentrifugation Facility.

RESULTS

A consensus TIM. The consensus sequence of all TIMs was determined from the most common amino acid in each position of the Pfam alignment (version 18.04) of 639 sequences. Because Hidden Markov Model alignment is not well suited to deal with insertions relative to the seed alignment, the total number of positions in the alignment (373) is much larger than the average length of a TIM sequence (235 aligned positions). Consequently, only positions with greater than 45% occupancy were selected, resulting in a sequence of 248 amino acids including four unaligned N- and C-terminal residues from *S. cerevisiae* TIM. (*S.c.* TIM is also 248 amino acids.) Because of the great evolutionary diversity of this ancient enzyme family, the consensus amino acid sequence is only 70% identical to that of *T. molitor* TIM, its closest known homolog.

The gene for the cTIM was assembled from synthetic oligonucleotides using a PCR scheme similar to the reassembly step in DNA shuffling (41). The gene was cloned into two expression vectors, one under the control of the *tac* promoter, and one under the control of the T7 promoter. The *tac* construct was transformed into DF502, an *E. coli* strain deficient in TIM and several other genes nearby in the chromosome (42). Growth on lactate and glycerol minimal media was comparable to complementation with *S.c.* TIM using the same construct. However, DF502 growth was inconsistent in our hands, perhaps because of the very slow growth on minimal media due to the large number of metabolic genes knocked out in this strain. We turned to the recent Keio collection single-gene knockout of TIM (38), which we lysogenized with DE3 phage to support transcription from the T7 promoter . At 5 □M IPTG, cTIM supported growth on

lactate minimal media in 2-3 days and on glycerol minimal media in 4 days, while *S.c.* TIM resulted in growth in about 1 day on both media.

The cTIM protein could be overexpressed at very high levels in *E. coli*, and was purified to near-homogeneity using two-step IMAC purification with 6×His tag cleavage by TEV protease. To eliminate contamination by the endogenous *E. coli* TIM, the engineered TIMs were purified from the Keio TIM-knockout DE3 strain. The Michaelis-Menten parameters were determined from steady-state kinetics for both directions of the isomerization reaction. The K_m values for DHAP and GAP are comparable to *S. cerevisiae* TIM, but the k_{cat} values are reduced by about 10^4-fold (*Table 1*). Wild-type TIMs exhibit bimolecular kinetics close to the diffusion limit, but apparently weak growth can be supported with significant reductions in activity. Therefore, an active TIM was derived from consensus alone, albeit one with significantly reduced activity.

Far-UV circular dichroism spectra for cTIM and *S. cerevisiae* TIM are similar and consistent with similar $(\square\square\square)_8$ architecture (*Fig. 1A-B*). Thermal denaturation was followed by CD spectroscopy at 222 nm. *S.c.* TIM unfolds in a single, irreversible step at about 60 °C. cTIM exhibits a similar pre-transition baseline to *S.c.* TIM, but does not unfold in a single step and is only ~50% unfolded at 95 °C (*Fig. 1C*). Unlike *S.c.* TIM, which precipitates at 95 ºC, cTIM shows no signs of precipitation and exhibits some reversibility on cooling from 95 ºC. This behavior is consistent with the thermal stabilization that has been observed for consensus mutations, although it is possible that more 'molten globule' character is also exhibited by cTIM.

With the exception of a few tetrameric TIMs from thermophiles, all known TIMs are homodimeric. The structure of TIM suggests that dimerization is necessary for full assembly of the active site by the interdigitation of loop 3 from the opposite monomer, and engineered monomeric TIMs exhibit k_{cat}/K_m values reduced by about 10^4-fold (43-46). The quaternary

structure of cTIM was determined by gel filtration chromatography (*Fig. 1D*). cTIM elutes significantly after *S.c.* TIM. Elution volumes were compared to a standard curve to determine apparent molecular weights; *S.c.* TIM eluted as the expected dimer (~56 kD), but the consensus enzyme elutes as a monomer at room temperature with an apparent molecular weight of ~29 kD. Surprisingly, the consensus sequence of over 600 dimeric proteins is a monomer.

Engineering the interface of cTIM. Although the monomeric state of cTIM was a surprise, its activity is consistent with TIM variants intentionally engineered to be monomers (43-46). These attempts to monomerize TIM involved deletions in the interfacial loop 3 and mutations that reversed charge pairing. We hypothesized that by choosing the most common amino acid at each position of cTIM we had scrambled necessary amino acid interactions (i.e., correlations) at the dimer interface. To examine this hypothesis, we reverted the dimeric interface to the sequence observed in *S.c.* TIM, which is known to be dimeric. The 1YPI crystal structure reveals 40 residues within 5 Å of the opposite momomer. The twelve interface residues that differed between cTIM and *S.c.* TIM were mutated in cTIM to create an interface reversion-cTIM (ir-cTIM, *Fig. 2A*).

The ir-cTIM was purified in similar yield to the original consensus TIM. Circular dichroism spectra are similar, but ir-cTIM exhibits greater signal at 205 nm suggesting more random coil. The thermal melts monitored at 222 nm were essentially identical (*Supplemental Fig. 3*). By gel filtration chromatography, ir-cTIM elutes at a calculated molecular weight slightly larger than cTIM at room temperature (~42 kD, *Fig. 2B*). Sedimentation velocity by analytical ultracentrifugation confirmed that the protein is still monomeric at room temperature (*Fig. 2C*). Furthermore, ir-cTIM did not exhibit concentration-dependant oligomerization over a ten-fold

range of concentrations (0.15-1.5 mg/mL). The activity of ir-cTIM was decreased compared to cTIM and failed to complement the Keio TIM knockout on minimal media.

When the gel filtration chromatography was repeated at 4 °C (*Fig. 2B*), all three of the proteins (*S.c.* TIM, cTIM and ir-cTIM) eluted as a dimer. For cTIM, a shoulder on the dimer-weight peak suggests that both monomer and dimer are populated at 4 °C and 37 □M (1 mg mL^{-1}), suggesting this is close to the K_D at this temperature. These results together suggest that the monomeric states of cTIM and ir-cTIM at room temperature are not a result of inherent defects in the dimerization interface, but rather non-native global dynamic properties of the cTIM scaffold. We analyzed the binding of the three proteins to the hydrophobic dye 1-analinonapthalene-8-sulfonic acid (ANS). ANS is quenched in aqueous buffer, but fluoresces strongly in lower dielectric environments such as organic solvent or when bound in the core of a protein. ANS binding is taken to be a sign of fluid tertiary structure exhibited by molten globules (47). *S.c.* TIM shows a weak fluorescence emission peak at 418 nm, but both cTIM and ir-cTIM have strong red-shifted fluorescence with peaks at 460 nm (*Fig. 2D*). The 600 MHz $^{1}H,^{15}N$-HSQC NMR spectrum of cTIM, however, displays a fair amount of amide peak dispersion for a protein of this size (*Supplemental Fig. 10*). Taken together, the biophysical data suggest that cTIM is monomeric and not as well folded at native TIMs at room temperature and above.

Concentration and temperature studies. To further examine the weak activity of cTIM, single point kinetics were observed over a range of enzyme concentrations at 4 °C and 37 °C (*Supplemental Fig. 9*). *S.c.* TIM, which is dimeric at both temperatures across the whole range of concentrations, increased in activity linearly with respect to concentration at both temperatures. Furthermore, there was a 13-fold decrease in activity at each concentration when

the reaction was performed at 4 °C versus 37 °C. When cTIM was assayed under the same conditions (at 60-240 □M enzyme, close to the apparent K_D at 4 °C), we still observed a linear increase in activity with respect to concentration at both temperatures, and the activity was 80-fold lower at the lower temperature for all three concentrations. If activity required dimerization, we would have expected a non-linear increase in activity at increasing concentration, as more of the dimeric state is populated. And we would have expected a smaller decrease in activity between 37 °C and 4 °C at all concentrations, since cTIM goes from mostly monomeric to mostly dimeric under these conditions. The composite data suggest that cTIM is active as a molten-globular monomer.

Database curation. A third consensus TIM variant that we engineered unexpectedly shed light on the properties of the original cTIM. When we began the analysis for correlated occurrences of amino acids, we downloaded the then-current version (22.0) of the Pfam database and curated it to remove repeated sequences and sequence fragments that did not represent full genes. More precisely, sequences with fewer than 205 aa and exact sequence repeats were removed from the 1,239 sequence database to yield 781 non-redundant full-length sequences (*Supplemental Fig. 1*). A new curated consensus TIM (ccTIM) was created using a similar approach to occupancy as described for cTIM, resulting in a 248 aa sequence with 36 sequence differences from cTIM (34 substitutions, an insertion and a deletion). These differences arise from changes to the database trivially affecting which amino acid is the most common in unconserved positions. The amino acid bias of a position can be quantified by calculating the relative entropy between positional distribution and the distribution of amino acids in a neutral reference state, such as amino acid usage in all open reading frames in yeast. From this calculation, it is evident that only unbiased positions were affected (*Fig. 3A*). These positions

tolerate virtually any amino acid in all TIMs, and therefore only minor differences were anticipated between cTIM and ccTIM.

A curated consensus TIM. ccTIM expresses well in bacteria with yields approaching 50 mg L^{-1}. CD wavelength spectra and thermal melt traces were essentially the same as cTIM (*Supplemental Fig. 3*). However, other biophysical properties turned out to be starkly different. When the thermal melt is reversed from 95 to 25 °C, ccTIM re-folds almost quantitatively. There is a red-shift in dye emission upon binding to ANS, but the very low level of fluorescence suggests that ccTIM is much less molten than cTIM (*Fig. 3C*). The protein elutes from a gel filtration column at room temperature with an apparent molecular weight of 66 kD, slightly more than *S.c.* TIM or the calculated dimeric mass (*Fig. 3D*). AUC sedimentation velocity studies confirm the protein is dimeric (50.5 kD with 95% confidence) with less than 2% forming higher aggregates (*Fig. 3E*).

ccTIM is nearly as active as wild-type TIMs, with comparable DHAP and GAP K_m values and k_{cat} values of 10^4-10^5 min^{-1}. ccTIM complements growth in the Keio TIM knockout, leading to growth on minimal media similar to *S.c.* TIM and faster than cTIM. Surprisingly, although cTIM and ccTIM differ only in a small number of unconserved positions and have similar structural and thermodynamic properties, cTIM is a molten globular monomer with weak activity and ccTIM is a native-like structured dimer with wild-type activity.

A comparison between consensus TIMs. While cTIM is 70% identical to *T. molitor* TIM, ccTIM is only 61% identical to its nearest natural sequence neighbor, *Roeiflexus sp* TIM. Only one mutated residue is within 5 Å of the active site residues (K12, H95, E165), the active site lid (residues 166-176), or the 2PG inhibitor bound in crystal structure 2YPI. The only proximal mutated position (I127V) is close by virtue of a backbone-backbone interaction with E165. The

mutations are spread throughout the protein secondary structures (17 helix, 10 sheet and 8 loop), and they are mainly solvent exposed (24 are more than 10% exposed with average exposure of 21%, *Fig. 3A*) (48). Except for F224A, the eight non-conservative mutations were on the protein surface (49). Stated simply, there is no obvious reason for the dramatic differences between the properties of cTIM and ccTIM.

The 36 differences between the consensus TIMs are at largely unconserved positions. The average relative entropy compared to the neutral reference state is 1.42 for all positions versus 0.82 for the 36 varying positions. Most of the twelve positions with relative entropies greater than 1.00 arise from distributions with a significant number of sequences occupied by two or three amino acids. For example, position 238 has a relative entropy of 1.38. The initial distribution was 169 Pro and 137 Ala out of 407 sequences occupied at this site. The curated distribution changed to 221 Pro and 325 Ala out of 720, switching the most common and next most common residues. A large fraction of the positions (eleven) were mutated to Ala, but the overall result was that the net charge of ccTIM is quite high (-11 in the 240 aligned positions, versus -5.5 for cTIM, -3.5 for *S.c.* TIM, and -5 ± 5 for TIMs overall). This phenomenon was seen before with the consensus sequence of the TPR motif, where it was shown to arise from scrambling of correlated surface charges (23).

We speculate that one major difference between cTIM and ccTIM may be in the extent of correlated occurrences of amino acids that are scrambled or broken. The majority of cTIM sequences are from eukaryotes, with 35% from the metazoans, while the majority of ccTIM sequences are bacterial with only 15% arising from the metazoans. Preliminary results from a mutual information analysis of the TIMs suggests that there is an extensive network of correlated residues in the metazoans, while correlated positions in bacteria are sparser and less well

connected (VD, BJS and TJM, manuscript in preparation). This suggests that the scrambling effect of the consensus sequence on correlated positions will be more detrimental to sequences with significant metazoan influence. We are making many further mutations to directly test this hypothesis as well as the roles of the most significant mutations between cTIM and ccTIM in terms of conservation, chemical dissimilarity, proximity to the active site and dimerization interface, and solvent exposure.

DISCUSSION

One clear and surprising lesson from this work is that an enzyme with native-like activity can be engineered from consensus alone, even for a large family of enzymes with significant evolutionary diversity that carry out a sophisticated and highly-tuned reaction. Natural TIMs exhibit nearly diffusion-controlled kinetics, which are believed to arise from a highly-orchestrated cycle of loop motion and precise positioning of residues in the active site to stabilize the enediol intermediate and avoid the formation of a toxic methylglyoxal byproduct. ccTIM is able to carry out this reaction at wild-type rates despite differing from the nearest natural TIM in 40% of its amino acids and never having been subject itself to evolution. This strongly argues that the vast majority of information for protein structure and function is encoded positionally, at the level of consensus, and not in higher-order correlations. (It would be interesting in the future to examine methylglyoxal formation by the TIM variants engineered here.)

However, the stark differences between ccTIM, cTIM and ir-cTIM illustrate that there is more information in the sequence families than just the positional information. These proteins are all in a sense “consensus” variants. They differ in sites that are highly tolerant to mutation, and they arise from variations between the most common amino acids at those unconserved positions. There is no obvious reason why the particular set of amino acids at the 36 positions

that differ between cTIM and ccTIM results in a weakly-active monomer in the former case and a wild-type-like dimer in the latter. The only striking difference between the two proteins at the sequence level is the fraction of eukaryotic (and especially metazoan) sequences that compose the cTIM and ccTIM databases. Our preliminary analysis suggests that an extensive network of correlated residues is present in the metazoans, and the scrambling of that network may be involved in the differences between the two consensus variants. (A complete analysis of TIM family correlation and how it relates to these two variants will be presented separately.) We are interested in engineering consensus versions of the metazoan, eukaryotic and prokaryotic enzymes to test this hypothesis directly, as well as in completing and "breaking" these correlated networks in cTIM and ccTIM.

While a native-like protein resulted from the curated database and a less-active molten globular protein resulted from the uncurated one, this does not necessarily suggest that curation is the key to successful consensus engineering. The sequence collections that are available significantly undersample complete evolutionary history and are affected by researcher interest and organism availability. For consensus design, it is difficult to articulate a convincing reason why any one sequence (or even sequence fragment) should be included or omitted from a sequence library, since the process by which the library was created was inherently biased. A related factor that we completely neglect here is that sequence alignment quality is likely to have a significant effect, especially on weakly conserved positions. Weakly conserved stretches and regions with length heterogeneity (such as loops) are the most difficult to align with certitude. Larger numbers of sequences certainly do improve alignment quality, and further expansion of sequence databases will likely improve our understanding of weakly-conserved positions and correlations among them.

The biophysical differences between cTIM and ccTIM are especially fascinating. Because of the way that the enzymes are designed, all of the conserved residues required for function (e.g., the Glu, His and Lys in the active site) are present. The consensus enzymes exhibit very similar CD spectra to yeast TIM, and even the weak activity of cTIM suggests that the proteins exhibit or at least sample highly similar structures to the natural TIMs. However, the oligomeric states and ANS binding data suggest that the primary difference between cTIM and ccTIM is in their global dynamic properties; that is, cTIM is more fluid and only dimerizes significantly at low temperature. It is still unclear how evolutionarily-plausible mutations at 36 unconserved positions result in this difference. Structural studies on ccTIM and covalent modification studies of cTIM are underway to understand better the nature of this dynamic shift.

While it is difficult to prove beyond a shadow of a doubt, the preponderance of the evidence argues that cTIM is active as a monomer. The most convincing evidence is that cTIM activity increases in direct proportion to concentration (implying that any additional dimerization is not increasing activity), and that cTIM actually takes a larger hit in activity than *S.c.* TIM upon cooling to 4 °C, even though *S.c.* TIM is a dimer at both concentrations and cTIM is significantly dimeric only at 4 °C. Further purification of cTIM by ion exchange chromatography did not result in higher activity, and multiple preparations yielded similar activities, suggesting that the problem is not simply that there is a large inactive population. Careful controls, including purification from a TIM-free strain, ensure that wild-type TIM contamination is not the cause of the activity.

Wierenga and colleagues have engineered several versions of trypanosomal TIM to be monomeric, and it has actually proven surprisingly difficult (43-46). Even relatively radical mutations or deletions to the interfacial loop 3 resulted in significant amounts of dimer at higher

concentrations. We believe that our concentration and temperature-dependent kinetics provides some of the strongest evidence that TIM can function as a monomer. However, it is interesting both that the trypanosomal monomeric mutants have similar k_{cat} values to cTIM and that the mechanism of monomerization is so different in cTIM/ir-cTIM (i.e., global scaffold verse interface mutations).

The unusual dynamic nature of cTIM calls to mind the loop motions present in the TIM catalytic cycle (30-35). Movement of loop 6 occurs on the same time scale as catalysis. As it appears to form a lid on the active site, its motion is thought to be coordinated with catalysis. This loop motion has been observed directly by fluorescence, solution and solid state NMR. One possibility is that cTIM's low activity is due in part to dysregulation of the loop motions. We attempted to make single Trp mutants of cTIM for ^{19}F-Trp incorporation and NMR studies analogous to those of McDermott et al., but the single Trp168 mutant (W11F W157F W191F) of cTIM is inactive. Further experiments to probe this issue in cTIM and ccTIM are underway.

Finally, it is a surprise that cTIM is even weakly active given its fluid nature, because the TIM reaction is thought to result from highly precise positioning of catalytic residues. The result is reminiscent of the recent discovery of Hilvert and colleagues that an engineered monomeric chorismate mutase from *Methanococcus jannaschii* (mMjCM) has similar catalytic efficiency to its native-like dimeric counterpart (50, 51). The balance of enthalpy and entropy changes upon substrate binding was dramatically altered for mMjCM, but with little net effect on the overall free energy. It will be interesting to calorimetrically analyze the binding of cTIM to inhibitors.

ACKNOWLEDGEMENTS

BJS was an NIH CBIP Fellow and Ohio State Presidential Fellow. We are grateful to Deepti Mathur for technical assistance with some of the enzyme preparation and kinetics. We thank

Christopher Jarnoiec and Jeffrey Lary for their expertise in NMR and AUC, respectively. This work was supported by The Ohio State University.

REFERENCES

1. Anfinsen CB (1973) Principles that govern the folding of protein chains. *Science* 181:223-230.
2. Kuhlman B, *et al.* (2003) Design of a novel globular protein fold with atomic-level accuracy. *Science* 302:1364-1368.
3. Dahiyat BI & Mayo SL (1997) De novo protein design: fully automated sequence selection. *Science* 278:82-87.
4. Cordes MH, Davidson AR, & Sauer RT (1996) Sequence space, folding and protein design. *Curr Opin Struct Biol* 6:3-10.
5. Richards FM (1997) Protein stability: still and unsolved problem. *Cell Mol Life Sci* 53:790-802.
6. Dill KA (1990) Dominant forces in protein folding. *Biochemistry* 29:7133-7155.
7. Rose GD & Wolfenden R (1993) Hydrogen bonding, hydrophobicity, packing, and protein folding. *Annu Rev Biophys Biomol Struct* 22:381-415.
8. Jiang L, *et al.* (2008) De novo computational design of retro-aldol enzymes. *Science* 319:1387-1391.
9. Rothlisberger D, *et al.* (2008) Kemp elimination catalysts by computational enzyme design. *Nature* 453:190-195.
10. Siegel JB, *et al.* (2010) Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction. *Science* 329:309-313.
11. Pabo C (1983) Molecular technology. Designing proteins and peptides. *Nature* 301:200.
12. Main ER, *et al.* (2003) Design of stable alpha-helical arrays from an idealized TPR motif. *Structure* 11:497-508.
13. Mosavi LK, Minor DL, Jr., & Peng ZY (2002) Consensus-derived structural determinants of the ankyrin repeat motif. *Proc Natl Acad Sci USA* 99:16029-16034.
14. Binz HK, *et al.* (2003) Designing repeat proteins: well-expressed, soluble and stable proteins from combinatorial libraries of consensus ankyrin repeat proteins. *J Mol Biol* 332:489-503.
15. Lehmann M, *et al.* (2000) From DNA sequence to improved functionality: using protein sequence comparisons to rapidly design a thermostable consensus phytase. *Protein Eng* 13:49-57.
16. Lehmann M, *et al.* (2002) The consensus concept for thermostability engineering of proteins: further proof of concept. *Protein Eng* 15:403-411.
17. Lehmann M, Pasamontes L, Lassen SF, & Wyss M (2000) The consensus concept for thermostability engineering of proteins. *Biochim Biophys Acta* 1543:408-415.
18. Steipe B, Schiller B, Pluckthun A, & Steinbacher S (1994) Sequence statistics reliably predict stabilizing mutations in a protein domain. *J Mol Biol* 240:188-192.
19. Ohage E & Steipe B (1999) Intrabody construction and expression. I. The critical role of VL domain stability. *J Mol Biol* 291:1119-1128.
20. Knappik A, *et al.* (2000) Fully synthetic human combinatorial antibody libraries (HuCAL) based on modular consensus frameworks and CDRs randomized with trinucleotides. *J Mol Biol* 296:57-86.

21. Godoy-Ruiz R, Perez-Jimenez R, Ibarra-Molero B, & Sanchez-Ruiz JM (2005) A stability pattern of protein hydrophobic mutations that reflects evolutionary structural optimization. *Biophys J* 89:3320-3331.
22. Pey AL, *et al.* (2008) Engineering proteins with tunable thermodynamic and kinetic stabilities. *Proteins* 71:165-174.
23. Magliery TJ & Regan L (2004) Beyond consensus: statistical free energies reveal hidden interactions in the design of a TPR motif. *J Mol Biol* 343:731-745.
24. Russ WP, *et al.* (2005) Natural-like function in artificial WW domains. *Nature* 437:579-583.
25. Socolich M, *et al.* (2005) Evolutionary information for specifying a protein fold. *Nature* 437:512-518.
26. Silverman JA, Balakrishnan R, & Harbury PB (2001) Reverse engineering the (beta/alpha)8 barrel fold. *Proc Natl Acad Sci USA* 98:3092-3097.
27. Alber T, *et al.* (1981) On the three-dimensional structure and catalytic mechanism of triose phosphate isomerase. *Philos Trans R Soc Lond B Biol Sci* 293:159-171.
28. Nickbarg EB & Knowles JR (1988) Triosephosphate isomerase: energetics of the reaction catalyzed by the yeast enzyme expressed in Escherichia coli. *Biochemistry* 27:5939-5947.
29. Nagano N, Orengo CA, & Thornton JM (2002) One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *J Mol Biol* 321:741-765.
30. Desamero R, *et al.* (2003) Active site loop motion in triosephosphate isomerase: T-jump relaxation spectroscopy of thermal activation. *Biochemistry* 42:2941-2951.
31. Rozovsky S, Jogl G, Tong L, & McDermott AE (2001) Solution-state NMR investigations of triosephosphate isomerase active site loop motion: ligand release in relation to active site loop dynamics. *J Mol Biol* 310:271-280.
32. Rozovsky S & McDermott AE (2001) The time scale of the catalytic loop motion in triosephosphate isomerase. *J Mol Biol* 310:259-270.
33. Williams JC & McDermott AE (1995) Dynamics of the flexible loop of triosephosphate isomerase: the loop motion is not ligand gated. *Biochemistry* 34:8309-8319.
34. Kempf JG, *et al.* (2007) Dynamic requirements for a functional protein hinge. *J Mol Biol* 368:131-149.
35. Wang Y, Berlow RB, & Loria JP (2009) Role of loop-loop interactions in coordinating motions and enzymatic function in triosephosphate isomerase. *Biochemistry* 48:4548-4556.
36. Gerlt JA & Raushel FM (2003) Evolution of function in (beta/alpha)8-barrel enzymes. *Curr Opin Chem Biol* 7:252-264.
37. Finn RD, *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res* 38:D211-222.
38. Baba T, *et al.* (2006) Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol* 2:2006 0008.
39. Plaut B & Knowles JR (1972) pH-dependence of the triose phosphate isomerase reaction. *Biochem J* 129:311-320.
40. Go MK, Koudelka A, Amyes TL, & Richard JP (2010) Role of Lys-12 in catalysis by triosephosphate isomerase: a two-part substrate approach. *Biochemistry* 49:5377-5389.

41. Stemmer WP, *et al.* (1995) Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides. *Gene* 164:49-53.
42. Babul J (1978) Phosphofructokinases from Escherichia coli. Purification and characterization of the nonallosteric isozyme. *J Biol Chem* 253:4350-4355.
43. Schliebs W, Thanki N, Jaenicke R, & Wierenga RK (1997) A double mutation at the tip of the dimer interface loop of triosephosphate isomerase generates active monomers with reduced stability. *Biochemistry* 36:9655-9662.
44. Borchert TV, Abagyan R, Jaenicke R, & Wierenga RK (1994) Design, creation, and characterization of a stable, monomeric triosephosphate isomerase. *Proc Natl Acad Sci USA* 91:1515-1518.
45. Borchert TV, *et al.* (1995) An interface point-mutation variant of triosephosphate isomerase is compactly folded and monomeric at low protein concentrations. *FEBS Lett* 367:315-318.
46. Borchert TV, *et al.* (1993) Overexpression of trypanosomal triosephosphate isomerase in Escherichia coli and characterisation of a dimer-interface mutant. *Eur J Biochem* 211:703-710.
47. Ptitsyn OB, *et al.* (1990) Evidence for a molten globule state as a general intermediate in protein folding. *FEBS Lett* 262:20-24.
48. Koradi R, Billeter M, & Wuthrich K (1996) MOLMOL: a program for display and analysis of macromolecular structures. *J Mol Graph* 14:51-55, 29-32.
49. Henikoff S & Henikoff JG (1992) Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA* 89:10915-10919.
50. Pervushin K, Vamvaca K, Vogeli B, & Hilvert D (2007) Structure and dynamics of a molten globular enzyme. *Nat Struct Mol Biol* 14:1202-1206.
51. Vamvaca K, *et al.* (2004) An enzymatic molten globule: efficient coupling of folding and catalysis. *Proc Natl Acad Sci USA* 101:12860-12864.

FIGURE LEGENDS

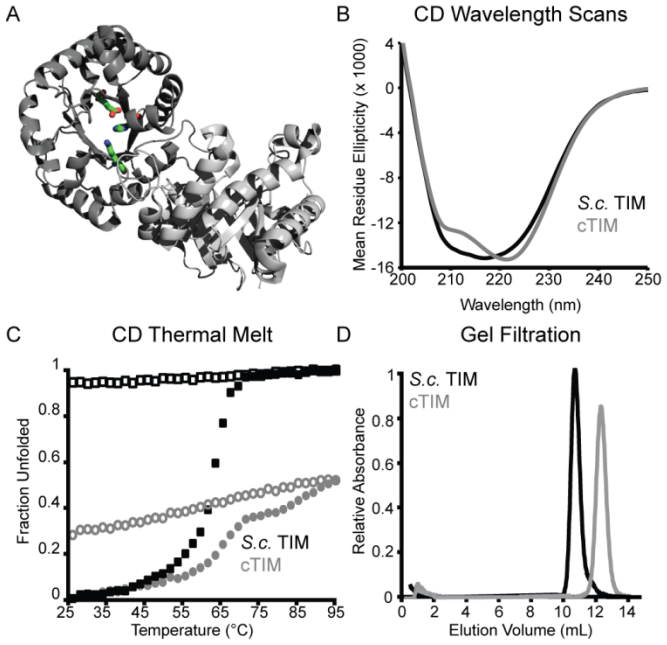

Fig. 1. cTIM characterization. (*A*) Crystal structure of *S. cerevisiae* TIM (1YPI) with active site residues shown. (*B*) CD wavelength scan of cTIM and *S.c.* TIM. (*C*) Thermal melt and cooling of cTIM and *S. c* TIM. Data collected at increasing temperatures are shown as closed points while data points collected during the reverse melt are shown open. (*D*) Gel Filtration shows that *S.c.* TIM elutes at a dimer, but cTIM elutes later with calculated molecular weight of a monomeric TIM.

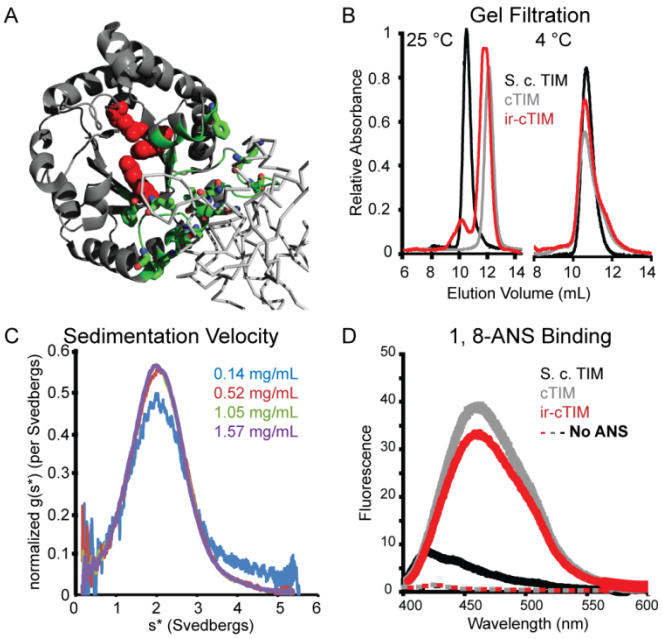

Fig. 2. ir-cTIM characterization. (*A*) Crystal structure of *S.c.* TIM with active site residues shown as red spheres. Interface mutations between cTIM and *S.c.* TIM depicted as green sticks and chain b is shown as a Cα-trace. (*B*) Elution volume from gel filtration of ir-cTIM calculates to a molecular weight intermediate of monomer and dimer. (*C*) Sedimentation velocity shows ir-cTIM is monomeric with no concentration-dependent oligomerization. (*D*) 1, 8-ANS binding of *S.c.* TM has a weak peak at 420 nm. cTIM and ir-cTIM exhibit strong fluorescence with a red-shift maxima of 460 nm.

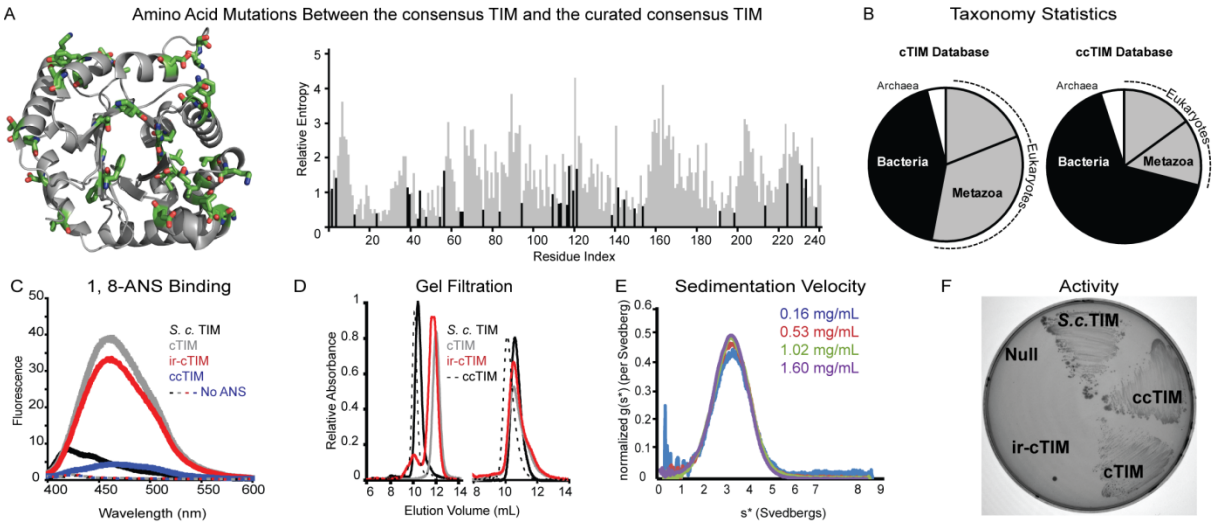

Fig. 3. ccTIM design and characterization. (*A*) The amino acid identities that change between the cTIM and ccTIM database are shown on the *S. c.* TIM structure. The relative entropies of all positions in the ccTIM dataset are plotted in gray with sites of mutation in black. (*B*) Distributions of taxonomies between datasets. (*C*). 1, 8-ANS binding of ccTIM yields a very weak maxima at 460 nm, but is non-fluorescent by eye. (*D*) ccTIM elutes slightly before the calculated volume for a dimer by gel filtration. (*E*) Sedimentation velocity confirms that ccTIM is dimeric with no concentration dependence between 0.16 and 1.6 mg mL^{-1}. (*F*) *In vivo* characterization of TIMs grown on minimal media lacking six-carbon sugars.